PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/66, 15/63, 15/13, 15/31, 15/38, C07H 21/04 — A1

(11) International Publication Number: WO 98/58067

(43) International Publication Date: 23 December 1998 (23.12.98)

(21) International Application Number: PCT/US98/12620

(22) International Filing Date: 18 June 1998 (18.06.98)

(30) Priority Data:
60/050,142 18 June 1997 (18.06.97) US

(71) Applicant *(for all designated States except US)*: EDGE BIOSYSTEMS, INC. [US/US]; 19212 Orbit Drive, Gaithersburg, MD 20879 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: SEED, John [US/US]; 9898 Carrigan Drive, Ellicot City, MD 21042 (US). SEED, Brian [US/US]; 9 Hawthorne Place #5J, Boston, MA 02114 (US). QIAN, Su, Wen [US/US]; 11531 Paramus Drive, North Potomac, MD 20878 (US). CANDAU, Reyes [US/US]; 3805 King William Drive, Olney, MD 20832 (US).

(74) Agents: KAGAN, Sarah, A. et al.; Banner & Witcoff, Ltd., 11th floor, 1001 G Street, N.W., Washington, DC 20001-4597 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: A VECTOR AND METHOD FOR PREPARATION OF DNA LIBRARIES

(57) Abstract

A method is presented for construction of a cDNA library in which the insert cDNA is flanked by linkers comprising recognition sites for an intron endonuclease. The insert may additionally comprise linkers comprising recognition sites for an endonuclease that generates non-self-complementary termini.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

A VECTOR AND METHOD FOR PREPARATION OF DNA LIBRARIES

Field of the Invention

The invention relates to the field of recombinant DNA. In particular, it relates to the field of DNA libraries.

Background to the Invention

The generation of high quality cDNA libraries is essential for gene identification strategies based on high throughput sequencing, on phenotypic expression in bacteria, yeast, or mammalian cells, on identification of interaction partners through the yeast two-hybrid system; or on recovery of cDNAs cognate to rare mRNAs. All of the varieties of expression cloning, for example, depend on the creation of, or access to, high quality cDNA libraries. Some of the key features of a high quality library include, a large number of independent clones (preferably greater than 10^7), a high percentage of inserts, oriented insertion of the genes, a high percentage of full length gene sequences and a population of clones which is representative of the starting population of RNAs. Many of the vectors and systems currently used for preparation of cDNA libraries have features which inherently limit the capacity to produce high quality libraries. To achieve the desired goal of orientation in a cDNA library, it is necessary to prepare cDNA with non-complementary ends. A typical strategy is to reverse transcribe the target RNA using an oligo-dT primer which comprises at its 5' end a restriction endonuclease recognition site. After preparing the complementary strand, adaptors are ligated onto the double stranded cDNA. These adaptors generate ends which are not complementary to the ends resulting from restriction of the site incorporated in the

5' end of the primer. The cDNA is then cut with this enzyme, purified to remove fragments and ligated into a suitably cleaved vector. To minimize loss of gene sequences, the enzymes used in this strategy (most commonly Not I) restrict DNA with low frequency. However, this enzyme is known to restrict DNA within coding elements and several important genes have been identified (from non-oriented libraries) that contain Not I restriction sites.

Another difficulty in the preparation of high quality libraries is background. For a variety of reasons, typical libraries have backgrounds (vector without insert) of 2 - 10%. Even though these backgrounds are sufficiently low for purposes such as expression cloning, they are unacceptable in high cost operations such as automated sequencing. Several strategies have been developed which reduce the incidence of background in cDNA libraries. The most common technique is to dephosphorylate the vector. A dephosphorylated vector cannot ligate to itself. By ligating a phosphorylated insert to a dephosphorylated vector, significant reductions in background can be obtained. However, vector dephosphorylation reduces ligation efficiency and thus library size. Furthermore, if the restriction enzyme used to cut the vector does not cut to completion, high backgrounds will result. In a variation on this strategy, the vector is cut with one restriction enzyme, dephosphorylated, then cut with two additional restriction enzymes to produce a phosphorylated vector with non-complementary ends which can then be used in high-efficiency ligations with phosphorylated inserts. However, this technique leads to ends which can still ligate to one another to produce vector dimers or insert dimers. The most efficient incorporation of cDNA is by methods that produce non self-complementary ends on both cDNA and vector. However, methods described to date do not allow directional insertion of cDNA. In yet another method to reduce background, toxic elements have been incorporated into vectors. These vectors are designed so that the cDNA insertion site resides within the promoter or coding elements of inducible genes which express products toxic to the host cell in the presence of an inducing agent. If cDNA is inserted in the vector, then the toxic gene will be interrupted and the host containing this plasmid will survive in the

presence of inducer whereas host cells containing plasmid vector without insert will die. However, the toxicity may be incomplete, resulting in a number of slow growing clones without insert. In addition, traces of nuclease activity can contribute substantially to background by changing the reading frame of restricted plasmid prior to ligation. Thus there is a need in the art for better methods of preparing oriented, representational cDNA libraries with very low backgrounds.

Detailed Description of the Invention

One aspect of this invention is a nucleotide polymer comprising two elements, one that binds to the nucleotide elements of another nucleic acid polymer and an element that comprises the recognition and cleavage site for an endonuclease. The binding elements include by way of example, but without limitation, random hexamers, random nanomers, homopolymers (deoxythymidine homopolymers in particular), sequence-specific nucleotides which bind to specific nucleic acid polymers, sequence-specific nucleotides which bind to known types or classes of nucleic acid polymers, and combinations of the above. The polymer of the invention is prepared by standard chemical or biological methods (Itakura, et al. 53 Ann. Rev. Biochem. 323-356 (1984); Current Protocols in Molecular Biology, Vol 1, Ausubel, et al., Eds, John Wiley & Sons, New York (1997)) known to those of skill in the art. The length of this nucleotide polymer element is variable and is dependent on conditions affecting the specificity of the binding and is typically between 4 and 200 nucleotides in length, with a preferred length between 6 and 30 nucleotides.

The element of the nucleotide polymer of the invention that comprises the recognition and cleavage site for an endonuclease is comprised of a nucleotide sequence that forms one complementary strand of a double stranded DNA sequence that is bound by an endonuclease and is cleaved by it. This nucleotide sequence is recognized only by endonucleases that cleave the coding elements of target genomes with a frequency less than 100 times per genome or by endonucleases

which do not cleave cDNA. The target genome is the DNA which comprises the source of the nucleic acid polymer to which the nucleotide polymer of the invention binds. An example of endonucleases that cleave the coding elements of target genomes with a frequency of less than 100 times per genome are the intron endonucleases. These endonucleases are intron-encoded enzymes that, under optimal conditions, recognize and cleave asymmetric DNA sequences of unusual length (14 - 31 bp). Intron endonucleases include by way of example, but without limitation, PI-Sce I (VDE), I-Ceu I, I-Tli I and I-Ppo I. Optimum conditions are those which are known to provide maximum specificity and enzyme activity. The preferred intron endonuclease of the invention is VDE. This enzyme has an unusually long asymmetric recognition and cleavage site, with only 1 known recognition and cleavage site in *S. cerevisiae*, and none in *E. coli*. By way of example, but without limitation, the wildtype recognition and cleavage sequence for this enzyme is 5'-TATGTCGGGTGCGGAGAAAGAGGTAAT GAAA-3' (Gimble and Wang, 263 J. Mol. Biol. 163-180 (1996)). It is known that this sequence is somewhat degenerate, i.e. base changes at certain locations enhance, decrease or have no impact on binding or cleavage. For example, substitutions of T for C at position -1 and G for C at position 6 generate a site which is more readily cleaved by VDE than the wild type. Other substitutions can be introduced at sites with known degeneracy to facilitate cleavage of the sequence, increase the GC content of the 3' overhang or introduce other changes as may be deemed valuable by those of skill in the art. Furthermore, substitution of Mn for Mg decreases the specificity of the enzyme. Consequently, changes to this sequence or to enzyme reaction conditions that preserve the functionality of the endonuclease (binding and cleavage) and that do not increase the frequency of cleavage to greater than 100 times per genome are within the scope of this invention.

Other enzymes whose recognition and cleavage sites are useful elements of the invention are those that recognize only modified nucleotides not normally found in cDNA or DNA amplified by *in vitro* technologies. One class of enzymes which meet this criterion are methylation-specific endonucleases. These enzymes

recognize only methylated DNA, and because cDNA or *in vitro* amplified DNA is not normally methylated (unless methyl nucleotides are deliberately introduced), it will not cleave these DNAs. The nucleotide sequence of the invention is methylated by chemical or enzymatic methods to produce a site which is cleaved by the methylation-specific endonuclease to produce a unique end. By way of example, but without limitation, a nucleotide sequence comprising overlapping Cla I sites can be methylated with Cla I methylase to produce a Dpn I recognition and cleavage site. Because this enzyme recognizes only methylated DNA, it will not cleave cDNA or *in vitro* amplified DNA. This example is provided only to illustrate the means by which those of skill in the art may identify and modify nucleotide sequences which serve as unique binding sites for endonucleases which recognize and cleave only modified nucleotides.

Other elements may be added to the nucleotide polymer of the invention as may be considered useful by those of skill in the art. These include by way of example, but without limitation, recognition and cleavage sites for other restriction endonucleases, sequences complementary to nucleotide polymers commonly used for sequencing, and recognition sites for DNA binding proteins. It is known in the art that the restriction and cleavage site for an endonuclease can be added to an existing DNA or RNA by techniques such as ligation of a double stranded oligonucleotide comprising the sequence for the restriction site, or by priming with a single stranded oligonucleotide comprising the sequence for the restriction site followed by extension of the first strand and synthesis of second strand, to create a duplex DNA which can be recognized and cleaved by the enzyme. The principle advantage of this invention is that it allows insertion of DNA into vectors without risk of cleaving the DNA which is being inserted. It is particularly useful in orienting DNA within vectors with little or no risk of cleaving the DNA which is being oriented.

It is an object of this invention to provide a method for inserting one DNA sequence within another DNA sequence wherein there is little or no risk of cleaving

the inserted DNA during subsequent manipulation of the DNA. In the method of the invention, adaptors are ligated to the inserted DNA before ligating the inserted DNA to the DNA sequence within which the DNA is to be inserted (vector). The vector comprises within the DNA insertion site one or more recognition and cleavage sites for an endonuclease which has less than 100 recognition and cleavage sites within the genome of the insert (rare endonuclease). The adaptors that are ligated to the insert DNA comprise either the full sequence of the recognition and cleavage site for the rare endonuclease of the invention, which must then be cleaved prior to insertion into the vector, or part of the recognition and cleavage site of the rare endonuclease which when ligated with the vector reconstitutes the full recognition and cleavage site for the rare endonuclease of the invention. The recognition and cleavage site for the rare endonuclease of the invention, if not already present, may be added to the vector using the same strategies as described for the insert DNA. When the insert of the method is ligated to the vector of the method, one or more recognition and cleavage sites for the rare endonuclease of the invention are regenerated.

It is another object of the invention to provide a vector with improved selection against clones lacking an insert. This vector has an element located between two polylinker sites which comprises a conditionally lethal gene. The region between the two polylinker sites also constitutes the cDNA insertion site. If the lethal gene is not removed during the process of vector preparation, host cells which are susceptible to the lethal gene product will be killed when transformed with this construct. In a preferred vector, the conditionally lethal gene is the *kilA* gene from the broad host range plasmid RK2. In host strains lacking the repressor genes *korA* and *korB*, the *kilA* gene is expressed and kills the host (Kornacki et al., 1993; Larsen and Figurski, 1994; Thomas et al., 1995; Thomson et al., 1993). On the other hand, vector DNA containing the *kilA* gene segment is prepared in good yield by standard methods from *E. coli* strains which constitutively express *korA* and *korB*. When the vector is transformed into a *korA, korB E. coli* strain, a normal transformation frequency (~10^9 colonies per ug DNA using

electrocompetent cells) is observed. The transformation frequency of *kor*⁻ bacteria, on the other hand, is between 1 and 10 colonies per ug DNA. This provides extremely powerful selection against vector.

It is yet another object of this invention to provide an improved vector for cloning of mammalian genes, expression of proteins in mammalian cells and expression cloning. In the vector of the invention, the CMV promoter of commonly used expression vectors is replaced with the more powerful and stable broad host range promoter from EF1α. The polylinker region of the vector is flanked by endonuclease sites which when cleaved, yield non-complementary ends. One of these ends is complementary to the cleavage product of an endonuclease that does not cleave cDNA. In a preferred form of the vector, the ends are complementary to the cleavage product of an intron endonuclease. The preferred endonuclease is VDE. Other sites downstream of the insertion site are provided to confer properties of improved stability and expression in transfected mammalian cells. These include the EBNA-1 transcription unit, SV40 and human growth hormone polyadenylation signal sequences, human IgG1 H/CH2 splice site, the Epstein-Barr virus *OriP*, puromycin acetyl transferase gene and transcription terminators. These preferred transcription termination sites are bidirectional, e.g. those from circular viral genome, such as those of the papova virus family, or synthetic bidirectional polyadenylation and termination signals (Figure 1).

It is an object of this invention to provide a method for the preparation of oriented plasmid cDNA libraries with greater than 10^8 primary transformants per ug of poly(A)+ RNA. Such libraries are prepared without the use of bacteriophage or bacteriophage vectors. The vector of this method is a plasmid with two or more endonuclease recognition and cleavage sites which when cleaved by one or more endonucleases, give noncomplementary ends. In a preferred vector, these sites are recognized by a single endonuclease with a degenerate cleavage site which can provide an overhang of 4 or more bases with two or more deoxyguanidines and/or deoxycytidines. The preferred site in this method is recognized and cleaved by Bst XI.

In the method of the invention, first strand synthesis is initiated with a nucleotide primer which binds to the target nucleic acid. These include, by way of example, but without limitation, random hexamers, random nanomers, homopolymers (deoxythymidine homopolymers in particular), sequence-specific nucleotides which bind to specific nucleic acid polymers, sequence-specific nucleotides which bind to known types or classes of nucleic acid polymers, and combinations of the above. Another element of the primer is the recognition and cleavage site for an endonuclease which is either not found in cDNA or found in very low frequency in target genomes. Examples of endonuclease recognition and cleavage sites that are not found in cDNA include by way of example but without limitation, the sites for intron endonucleases VDE, I-Ceu I, I-Tli I and I-Ppo I as well as for the methylation specific enzyme Dpn I. The preferred site is recognized and cleaved by the intron endonuclease VDE (Gimble and Stephens, 1995; Gimble and Thorner, 1992; Gimble and Thorner, 1993; Gimble and Wang, 1996). Cleavage of this site leaves a 4 base, 3' overhang that is rich in GC content but not palindromic (GTGC). This is an important feature in the preparation of large size cDNA libraries as it prevents end-to-end ligation of cDNA during ligation with vector. The high GC content of the overhang increases the stability of overlapping complementary ends. This in turn enhances ligation efficiency, which in turn

enhances library size. Substitutions of T for C at position -1 and G for C at position 6 generate a site which is more readily cleaved by VDE than the wild type. Other substitutions can be introduced at sites with known degeneracy to facilitate desired changes in the sequence. The primer of the invention is used to initiate reverse transcription. Many reverse transcriptases and some thermostable polymerases with reverse transcriptase activity have been described as being useful in the synthesis of first strand cDNA. First strand cDNA is synthesized by these or related enzymes according to standard procedures. This step is followed by second strand synthesis and ligation of phosphorylated, non-selfcomplementary adaptors. Second strand synthesis is performed using a suitable DNA polymerase such as T4 DNA polymerase or *E. coli* DNA polymerase I, and priming strategies such as RNase H treatment. Other enzymes, such as thermostable polymerases, and/or other priming strategies may be used which are known to those of skill in the art. Phosphorylated adaptors are selected, annealed and ligated to the cDNA essentially as described previously (Seed and Aruffo, 1987; Aruffo and Seed, 1987a). The key feature of this method is that the adaptors are non-selfcomplementary, i.e. annealing of the two strands of the adaptor generates an overhang which is not complementary to itself. By ligating the adaptors in large molar excess over the cDNA, end-to-end ligation of the cDNA is minimized. Although large amounts of end-to-end ligations of the adaptors occur, this is more than offset by efficient ligation of adaptors to the cDNA.

The endonuclease site (introduced through the primer) is then cleaved, leaving nonidentical, noncomplementary ends. It is preferred that the endonuclease not cleave the cDNA or that it cleave with very low frequency. The cDNA is fractionated and DNA greater than 0.5 - 1 kb in length is ligated to the vector and transformed into a suitable host. The preferred method of fractionation is potassium acetate gradients, although size exclusion chromatography, other density gradients or other techniques known to those of skill in the art may be used.

The preferred vector of the method has a cDNA insertion site which comprises a toxic stuffer gene, preferably the *kilA* gene of the invention, and two flanking restriction sites, cleavage of which leaves non-selfcomplementary overhangs that are complementary to those of the adaptor and the cleaved intron endonuclease site on the cDNA. The overhang which is complementary to the adaptor will be located at the 5' end of the cDNA in the preferred version. This strategy has the benefit of providing oriented inserts, and cDNA and vector which are completely non-selfcomplementary. The cDNA cannot ligate to itself, nor can the vector, in any orientation. As a result the cDNA is assimilated into the vector with maximal efficiency, in the correct orientation, with low background.

Other vector-based strategies for producing low background may be coupled with the intron endonuclease/non-self complementary adaptor strategy for producing large unbiased libraries. To achieve the large library sizes afforded by the method of the invention, such vectors must have phosphorylated, non-selfcomplementary ends which are complementary to the ends of the cDNA. A preferred enzyme for restriction of the preferred vector is Bst XI. This enzyme has a degenerate cleavage site which facilitates the selection of overhangs which are complementary to more than one restriction site and which may be manipulated to have other useful features such as high GC content. Thus, by engineering the correct sequences into different Bst XI restriction sites flanking the insertion site, a single restriction digest can generate ends which are complementary to both the adaptor and the cleaved VDE site. Other enzymes with degenerate cleavage sites known to those of skill in the art may be used to leave overhangs on the vector that are capable of annealing and ligating to the non-selfcomplementary adaptors described above and to the overhang generated by the intron endonuclease. Alternatively, more than one enzyme may be used to generate the appropriate vector configuration. It is generally advantageous, but not essential, that the overhang created by the adaptor is complementary to the overhang generated by the enzyme used to cleave the insertion site of the vector. However, the vector must be treated in such a way that non-selfcomplementary overhangs are created on the

vector which are complementary to the non-selfcomplementary overhangs on the cDNA. This can be achieved by ligating adaptors to appropriate restriction cleavage site of the vector or by similar techniques known to those of skill in the art. Alternative strategies which require the use of two enzymes or multiple manipulations of the vector are not preferred because they increase cost and the extra manipulations tend to reduce efficiency of library preparation.

For purposes of preparing a cDNA library, the plasmid vector containing the *kilA* gene is amplified in a *korA*, *korB* strain such as MC1061/p3 and isolated by standard methods. The purified plasmid is then digested with restriction enzymes that flank the *kilA* gene fragment. The *kilA* gene fragment is then separated from the vector by methods such as gel electrophoresis, size exclusion chromatography, density gradient centrifugation and other techniques known to those of skill in the art. The vector is then ready for further manipulations or immediate ligation of cDNA .

It is an object of this invention to provide a method for orienting DNA inserts within vectors with little or no risk of cleaving the inserted DNA. In this method, the same approach used to prepared oriented cDNA inserts is used genetically to orient DNA in plasmids and other DNA vectors used in the cloning and manipulation of DNA, with little or no risk of cleaving the inserted DNA.

Example 1. **Preparation of unamplifed cDNA library from human small intestine**

A cDNA library with 1.1×10^8 recombinant clones was prepared from human small intestine. The cDNA was prepared from 2.5 ug of poly(A)+ RNA using an oligo-dT-VDE primer comprising 18 bases of dT and 66 bases including 31 bases of the original VDE recognition site (underlined). The sequence of the primer is :

5'-CGACGTTGTAAAACGACGGCCAGTGAATTCTC
TATGTCGGGTGCGGAGAAAGAGG
TAATGAAA TACTTTTTTTTTTTTTTTTT-3'

Detailed protocol: The poly(A)RNA was diluted with DEPC-H_2O to 22 ul and mix with 2 ul of the primer (1 ug/ul). The mixture was incubated at 70°C for 10 minutes and transferred to ice for 5 minutes. The following components were added to the sample for the first strand cDNA synthesis: 8 ul of 5x 1st strand buffer [250mM Tris-HCl(pH 8.3), 375mM KCl, 15mM $MgCl_2$], 4 ul of 10 mM dNTP, 2 ul of 0.1 M DTT, 1 ul (40U) of RNase inhibitor (Life Technologies, Inc., Gaithersburg, MD) and 1 ul (36 U) of AMV RT-XL (Life Science Research Products, Orlando, FL). The sample was incubated at 42°C for 1 hr and 70°C for 10 minutes. The second strand cDNA was prepared by adding the following reagents to the first strand cDNA sample: 70 ul of H_2O, 30 ul of 5x 2nd strand Buffer [100 mM Tris-HCl (pH 6.9), 450 mM KCl, 23 mM $MgCl_2$, 0.75 mM β-NAD^+, 50 mM $(NH_4)_2SO_4$], 2 ul of 0.1 M DTT, 3 ul of 10 mM dNTP, 0.5 ul (2U) of RNase H (Life Technologies, Inc., Gaithersburg, MD), 4 ul (40U) of *E. coli* DNA polymerase I (Life Technologies, Inc., Gaithersburg, MD) and 1 ul (10U) of *E. coli* DNA ligase (Life Technologies, Inc., Gaithersburg, MD). The sample was incubated 2 hours at 16°C, then added 5 ul (5U) of T4 DNA polymerase (Boehringer Mannheim, Indianapolis, IN) and incubated 5 minutes at 16°C. The reaction was terminated by adding 10 ul of 0.5 M EDTA. The cDNA was purified by phenol extraction and ethanol precipitation. The ligation of the Bst XI adaptor to the cDNA was performed by adding the following reagents to 20 ul of the cDNA sample in H_2O: 10 ul of phosphorylated Bst XI adaptors (5'-CTGGCTCA-3'; 5'-TGAGCCAGCCCC-3') (10 ug), 10 ul of ligation buffer [330mM Tris-HCl, 50 mM MgCl2, 5 mM ATP], 7 ul of 0.1 M DTT and 5 ul (5U)of T4 DNA ligase (Life Technologies, Inc., Gaithersburg, MD) and incubating overnight at 16°C. The cDNA was purified by phenol extraction and ethanol precipitation, then digested with 20 units of VDE (New England Biolabs, Beverly, MA) at 37°C for 6 hours.

The cDNA was fractionated on a potassium acetate gradient (5 - 20%) for 3 hours at 50,000 rpm in a Beckman L5-50 centrifuge using an SW-50 rotor. The cDNA fragments greater than 700 bases were collected, concentrated by ethanol precipitation and dissolved in 50 ul of TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA). The pEAK8 vector was digested with Bst XI using manufacturer's recommended procedures (New England Biolabs, Beverly, MA) and purified on a potassium acetate gradient as above to remove the *kilA* stuffer. After ethanol precipitation, the vector was dissolved in TE at 50 ng/ul. Mix 47 ul of the vector with 47 ul of the fractionated cDNA , 258 ul of H_2O, 94 ul of 5x T4 DNA ligase buffer (Life Technologies, Inc., Gaithersburg, MD) and 24 ul of ligase (1 U/ul) (Life Technologies, Inc., Gaithersburg, MD). Incubate the ligation mix for 2 hours at room temperature. The DNA sample was desalted by ethanol precipitation and dissolved in 30 ul of H_2O. The DNA was electroporated into electrocompetent *E. coli* DH10B cells in 10 equal fractions, 0.3 ml per fraction. The transformed cells were incubated in SOC medium for 45 minutes before adding glycerol to 15% to make the frozen stock (-70°C) of the cDNA library. To check the library titer,10 ul of the library stock was diluted 200 times and plated on ampicillin LB-agar plates. The average size of the cDNA inserts was analyzed by extracting the plasmid DNA from 24 randomly selected colonies and digested with the restriction enzymes flanking the cDNA insert. In summary, the primary cDNA library for human small intestine contains 1.1x 10^8 primary transformants with the average size of the cDNA inserts at 2.3 kb. The vector (without insert) background in this library is less than 1%.

Example 2. **Preparation of unamplifed cDNA library from human fetal kidney**

A cDNA library with 2.7 x 10^6 recombinant clones was constructed from human fetal kidney RNA. The cDNA was prepared from 5.0 ug of poly(A)+ RNA using another oligo-dT-VDE primer comprising 18 bases of dT and 60 bases including 28 bases of the modified VDE recognition site (underlined). For example,

a variant with modification of two bases (C to T at position -1 and C to G at position 6 of the original version) and deletion of three deoxyadenines at the 3' of the original version leads to a site that can also be cleaved by VDE enzyme. The sequence of the primer is :

5'-CGACGTTGTAAAACGACGGCCAGTGAATTCT*T*
TATGTGGGGTGCGGAGAAAGAGG
TAATG TTTTTTTTTTTTTTTTT-3'

The cDNA library for human fetal kidney was prepared with similar protocol as above except the following changes: 1). the use of the modified oligo-dT-VDE primer ; 2). 2 hours of VDE digestion. This library contains 2.7 x 10^8 primary clones with average size of the cDNA inserts at 1.3 kb and less than 1 % of background.

Example 3. **High level protein expression in mammalian cells using pEAK10**

pEAK10 was prepared from pEAK8 by deleting an inhibitory regulatory sequence present in the EF1α promoter of pEAK8 and removing the BspLU 11I site in the EFIα promoter. The protein expression levels from pEAK10 are 50% higher than those for pEAK8.

The *LacZ* gene from *E. coli* was cloned in pEAK10 and in three other different commercial vectors and the resulting plasmids were transfected into 293HEK cells expressing the EBNA-1 protein and the large T-antigen (293 EBNA-T). The amount of recombinant protein expressed in 293 EBNA-T cells transfected with pEAK10 was, at least, three fold higher than when the same cells were transfected with any of the other plasmids.

LacZ was cloned by standard methods (Current Protocols in Molecular Biology, Vol 1, Ausubel, et al., Eds, John Wiley & Sons, New York (1997)) into the Hind III-Not I sites of pEAK10 or pCDNA3.1/Hygro (+) [(Invitrogen, cat#

V870-20) (this vector has the CMV promoter and the SV40 origin of replication)], pREP4 [(Invitrogen, cat# V004-50) (this vector has the RSV promoter, the EBNA-1 expression cassette and an Epstein-Barr virus origin of replication)] or pCEP4 [(Invitrogen, cat# V044-50) (this vector has the CMV promoter, the EBNA-1 expression cassette and an Epstein-Barr virus origin of replication)] to generate pEAK10-βgal, pCDNA3-βgal, pREP4-βgal or pCEP4-βgal.

5×10^5 293 EBNA-T cells were plated in 60 mm Petri dishes containing 5 ml DMEM medium supplemented with 10% calf serum and incubated at standard conditions (37°C and 5% CO_2) for 24 hours. The medium was then changed and the plates were incubated in the same conditions for two additional hours. 3 ug of pEAK10-βgal (or pCDNA3-βgal, or pREP4-βgal , or pCEP4-βgal, three samples of each plasmid) were pipetted into a microtube and the following components were added in the following order: up to 225 ul water, 25 ul 2.5 M $CaCl_2$, 250 ul [50mM HEPES, pH 7.05, 1.26 mM Na_2HPO_4, 140 mM NaCl], the mix was vortexed briefly, incubated at room temperature for one minute and then added dropwise to the cell cultures. After three hours of incubation at standard conditions, the medium was changed and the transfected cells were incubated for 1, 2 or 3 days. (A total of twelve experiment were done, result from the combination of four different plasmids at three expression times).

To harvest the recombinant protein (β-galactosidase), the cells were washed once with PBS and collected in 1 ml PBS, spun at 250 g for 5 minutes and resuspended in 100 ul 0.25 M Tris.ClH pH 8. The cells were then lysed by three freeze-thaw (liquid nitrogen/37°C water bath) cycles and the insoluble material was pelleted at 12000 g for 5 minutes at 4°C.

β-galactosidase liquid assays were done by standard protocols (Current Protocols in Molecular Biology, Vol 1, Ausubel, et al., Eds, John Wiley & Sons, New York (1997)), and samples were quantified against a standard curve using purified *E. coli* β-galactosidase (Sigma. St. Louis, MO).

The levels of β-galactosidase (expressed as % of β-galactosidase per total amount of protein) showed that pEAK10 was superior in protein expression to any of the other plasmids at any given time (Table I).

	pEAK10-βgal	pCDNA3-βgal	pREP4-βgal	pCEP4-βgal
day 1	0.55	0.45	0.21	0.35
day 2	0.87	0.59	0.20	0.40
day 3	105	0.50	0.31	0.48

TABLE I: Comparison of expression levels of β-galactosidase (shown as % of β-galactosidase relative to total protein) in pEAK10 versus other three commercial plasmids at different time points.

Literature Cited

Aruffo, A., and Seed, B. (1987a). Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. Proceedings of the National Academy of Sciences of the United States of America *84*, 8573-7.

Aruffo, A., and Seed, B. (1987). Molecular cloning of two CD7 (T-cell leukemia antigen) cDNAs by a COS cell expression system. EMBO Journal *6*, 3313-6.

Brenneman, M., Gimble, F. S., and Wilson, J. H. (1996). Stimulation of intrachromosomal homologous recombination in human cells by electroporation

with site-specific endonucleases. Proceedings of the National Academy of Sciences of the United States of America *93*, 3608-12.

Gimble, F. S., and Stephens, B. W. (1995). Substitutions in conserved dodecapeptide motifs that uncouple the DNA binding and DNA cleavage activities of PI-SceI endonuclease. Journal of Biological Chemistry *270*, 5849-56.

Gimble, F. S., and Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. Nature *357*, 301-6.

Gimble, F. S., and Thorner, J. (1993). Purification and characterization of VDE, a site-specific endonuclease from the yeast Saccharomyces cerevisiae. Journal of Biological Chemistry *268*, 21844-53.

Gimble, F. S., and Wang, J. (1996). Substrate recognition and induced DNA distortion by the PI-SceI endonuclease, an enzyme generated by protein splicing. Journal of Molecular Biology *263*, 163-80.

Kornacki, J. A., Chang, C. H., and Figurski, D. H. (1993). kil-kor regulon of promiscuous plasmid RK2: structure, products, and regulation of two operons that constitute the kilE locus. Journal of Bacteriology *175*, 5078-90.

Larsen, M. H., and Figurski, D. H. (1994). Structure, expression, and regulation of the kilC operon of promiscuous IncP alpha plasmids. Journal of Bacteriology *176*, 5022-32.

Seed, B., and Aruffo, A. (1987). Molecular cloning of the CD2 antigen, the T-cell erythrocyte receptor, by a rapid immunoselection procedure. Proceedings of the National Academy of Sciences of the United States of America *84*, 3365-9.

Thomas, C. M., Smith, C. A., Ibbotson, J. P., Johnston, L., and Wang, N. (1995). Evolution of the korA-oriV segment of promiscuous IncP plasmids. Microbiology *141*, 1201-10.

Thomson, V. J., Jovanovic, O. S., Pohlman, R. F., Chang, C. H., and Figurski, D. H. (1993). Structure, function, and regulation of the kilB locus of promiscuous plasmid RK2. Journal of Bacteriology *175*, 2423-35.

What is claimed is:

1. A method for inserting one DNA sequence within another DNA sequence with little or no risk of cleaving the inserted DNA, comprising the steps of:
   a. preparing a double stranded DNA sequence to be inserted (insert DNA) by ligation of a double stranded oligonucleotide which either comprises a sequence that is recognized and cleaved by an endonuclease for which less than 100 recognition and cleavage sites exist within the genome of the target or which when ligated with a vector reconstitutes a sequence that is recognized and cleaved by an endonuclease for which less than 100 recognition and cleavage sites exist within the genome of the target;
   b. cleaving the insert DNA with an enzyme which recognizes and cleaves the insert DNA at the rare endonuclease recognition and cleavage site provided with the primer;
   c. ligating the cleaved insert DNA to the DNA sequence within which it is to be inserted (vector DNA), wherein the ends of the vector DNA are complementary to the ends of the cleaved insert DNA.
2. A method for inserting one DNA sequence within another DNA sequence with little or no risk of cleaving the inserted DNA, comprising the steps of:
   a. preparing a double stranded DNA sequence to be inserted (insert DNA) from RNA or DNA (target) using a nucleotide polymer (primer) comprising two elements, an element at one end of the primer that is the site for initiation of polymerization of the complementary DNA strand and another element at the other end of the primer that comprises one strand of a double stranded DNA sequence that is recognized and cleaved by an endonuclease for which less than 100 recognition and cleavage sites exist within the genome of the target;

b. preparing a second strand of DNA using the first strand as a template;

c. cleaving the insert DNA with an enzyme which recognizes and cleaves the insert DNA at the rare endonuclease recognition and cleavage site provided with the primer;

d. ligating the cleaved insert DNA to the DNA sequence within which it is to be inserted (vector DNA) wherein the ends of the vector DNA are complementary to the ends of the cleaved insert DNA.

3. A method for orienting one DNA sequence within another DNA sequence with little or no risk of cleaving the inserted DNA, comprising the steps of:

a. preparing a double stranded DNA sequence to be inserted (insert DNA) from RNA or DNA (target) using a nucleotide polymer (primer) comprising two elements, an element at one end of the primer that is the site for initiation of polymerization of the complementary DNA strand and another element at the other end of the primer that comprises one strand of a double stranded DNA sequence that is recognized and cleaved by an endonuclease for which less than 100 recognition and cleavage sites exist within the genome of the target;

b. preparing a second strand of DNA using the first strand as a template;

c. cleaving the insert DNA with an enzyme which recognizes and cleaves the insert DNA at the rare endonuclease recognition and cleavage site provided with the primer, wherein the ends of the cleaved insert DNA are distinct and not self-complementary;

d. ligating the cleaved insert DNA to the DNA sequence within which it is to be inserted (vector DNA), wherein the ends of the vector DNA are distinct and not self-complementary and are complementary to the ends of the cleaved insert DNA.

4. The method of claim 1 in which one end of the insert DNA is prepared by ligating non self-complementary adaptors to the DNA before cleaving with

an enzyme which recognizes and cleaves the rare endonuclease recognition and cleavage site provided with the primer.

5. The method of claim 2 in which one end of the insert DNA is prepared by ligating non self-complementary adaptors to the DNA before cleaving with an enzyme which recognizes and cleaves the rare endonuclease recognition and cleavage site provided with the primer.

6. The method of claim 3 in which one end of the insert DNA is prepared by ligating non self-complementary adaptors to the DNA before cleaving with an enzyme which recognizes and cleaves the rare endonuclease recognition and cleavage site provided with the primer.

7. A synthetic nucleotide sequence comprising the recognition and cleavage site for an intron endonuclease.

8. The nucleotide sequence of claim 7 which comprises less than a whole genome.

9. A plasmid vector which contains an insertion site for DNA that when cleaved gives two distinct and non self-complementary single stranded DNA sequences.

10. The vector of claim 9 which contains a bacterial origin of replication and a gene conferring a drug resistance in bacteria.

11. The vector of claim 9 which contains a gene located in the DNA insertion site that confers a conditionally lethal phenotype.

12. The vector of claim 11 in which the conditionally lethal gene is *KilA*.

13. The vector of claim 10 in which an EF1α promoter is upstream the insertion site for the cDNA.

14. The vector of claim 13 in which the insertion site for the DNA is followed by a human IgG1 H/CH2 splice sequence and a human polyadenylation signal sequence.

15. The vector of claim 12 which is able to express puromycin acetyl transferase gene..

16. The vector of claim 12 which contains the EBNA-1 transcription unit, and an Epstein-Barr virus origin of replication.

17. The vector of claim 15 which contains an SV40 origin of replication.

18. The vector of claim 8 which contains a mammalian expression unit.

19. The vector of claim 18 that contains an origin of replication for mammalian cells.

20. A plasmid vector which contains the EF1α promoter, the EBNA-1 transcription unit, and an Epstein-Barr virus origin of replication.

21. A method for the preparation of oriented cDNA libraries comprising the steps of:

a. preparing DNA from RNA using a nucleotide polymer (primer) comprising two elements, an element at one end of the primer that is the site for initiation of polymerization of the complementary DNA strand and another element at the other end of the primer that comprises one strand of a double stranded DNA sequence that is recognized and cleaved by an endonuclease for which less than 100 recognition and cleavage sites exist within the genome of the target;

b. preparing a second strand of DNA using the first strand as a template;

c. cleaving the cDNA with an enzyme which recognizes and cleaves the cDNA at the rare endonuclease recognition and cleavage site provided with the primer, wherien the ends of the cDNA are distinct and non self-complementary;

d. ligating the cleaved cDNA is ligated to a plasmid vector, wherein the ends of the vector DNA are distinct and not self-complementary and are complementary to the ends of the cDNA.

22. The method of claim 21 in which one end of the cDNA is prepared by ligating non self-complementary adaptors to the cDNA before cleaving with an enzyme which recognizes and cleaves the rare endonuclease recognition and cleavage site provided with the primer.

23. The method of claim 22 in which the adaptors are phosphorylated.

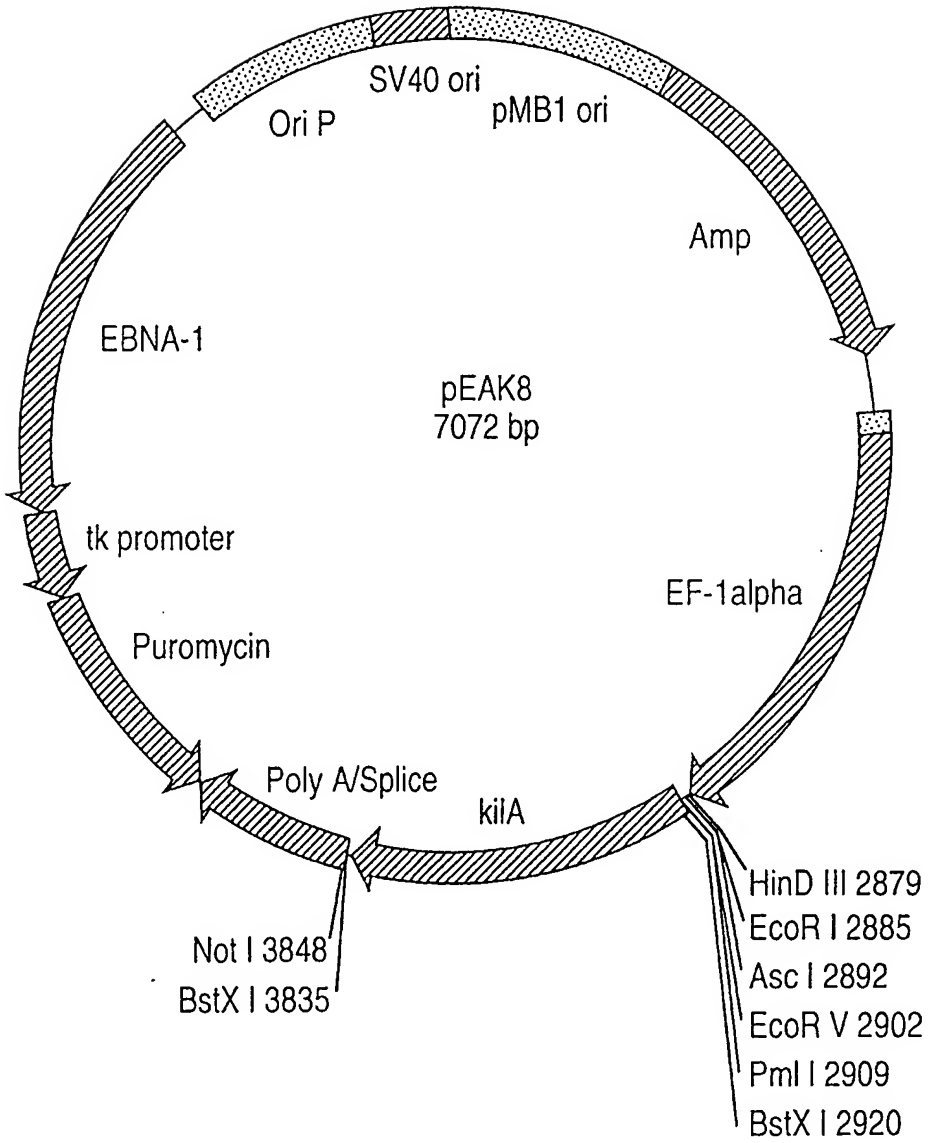
24. A method for production of proteins comprising the step of:

transfecting mammalian cells with a vector which comprises a EF1α promoter, an Epstein-Barr origin of replication, and an EBNA-1 transcription element.

25. The method of claim 24 wherein the mammalian cells express the T antigen of SV40 and the vector comprises the SV40 origin of replication.

26. The vector of claim 20 which comprises the SV40 origin of replication.

FIG. 1


SV40 ori
Ori P
pMB1 ori
Amp
EBNA-1
pEAK8
7072 bp
tk promoter
EF-1alpha
Puromycin
Poly A/Splice
kilA
HinD III 2879
EcoR I 2885
Asc I 2892
EcoR V 2902
PmI I 2909
BstX I 2920
Not I 3848
BstX I 3835

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/12620

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12N 15/66, 15/63, 15/13, 15/31, 15/38;C07H 21/04

US CL : 435/91.41, 320.1; 536/23.5, 23.7, 23.72, 24.1, 24.2

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/91.41, 320.1; 536/23.5, 23.7, 23.72, 24.1, 24.2

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, Medline, WPIDS, Biosis

search terms: intron endonuclease, kilA, EF1 alpha, IgG1, splice, puromycin acetyl transferase

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KLICKSTEIN et al. (a) Preparation of insert DNA from messenger RNA. In Current Protocols in Molecular Biology (Ausubel et al. Editors) John Wiley & Sons, New York. 1995, Pages 5.5.1-5.5.14, especially pages 5.5.6-5.5.9.	3, 6
Y	KILICKSTEIN et al. (b) Ligation of Linkers or Adapters to Double-Stranded cDNA. In Current Protocols in Molecular Biology (Ausubel et al. Editors) John Wiley & Sons, New York. 1991, pages 5.6.1-5.6.10, especially page 5.6.5.	1-8, 18, 21-23

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 02 SEPTEMBER 1998	Date of mailing of the international search report 14 OCT 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JOHN S. BRUSCA Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/12620

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y -- A	Invitrogen Catalog. Expression Systems and Vectors. San Diego, CA: Invitrogen Corporation 1994, pages 49-54, especially pages 49-54.	9, 10 --------- 18-20, 24-26 --------- 16, 17
Y	New England Biolabs Catalog. Beverly, MA: New England Biolabs, Inc., 1993/1994, page 52, especially page 52.	1-8, 18, 21-23
Y	LAMBOWITZ et al. Introns as mobile genetic elements. Annual Reviews in Biochemistry. 1993, Vol. 62, pages 587-623, especially pages 595-599.	1-8, 18, 21-23
A	VARA et al. Expression in mammalian cells of a gene from Streptomyces alboniger conferring promycin resistance. Nucleic Acids Research. 1986, Vol. 14, No. 11, pages 4617-4624, especially page 4617.	15, 17
Y	GILLIES et al. High-level expression of chimeric antibodies using adapted cDNA variable region cassettes. J. Immunological Meth. 1989, Vol. 125, pages 191-202, especially page 193.	14
Y	KIM et al. An efficient expression vector for stable expression in human liver cells. Gene. 1993, Vol. 134, pages 307-308, especially pages 307-308.	13, 20, 24-26
A	FIGURSKI et al. Broad host range plasmid RK2 encodes multiple kil genes potentially lethal to Escherichia coli cells. Proc. Natl. Acad. Sci. USA. March, 1982. Vol. 79, pages 1935-1939, especially page 1935.	11, 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/12620

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/12620

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group 1, claim(s)1-6, and 21-23, drawn to a method of inserting DNA into a vector.
Group 2, claim(s) 7, 8, 18, and 19, drawn to nucleic acid comprising an intron endonuclease cleavage site.
Group 3, claim(s) 9-17, 24, and 25, drawn to a plasmid vector and methods of using the plasmid vector to express a gene.
The inventions listed as Groups 1-3 do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: Rule 13.2 requires that the claimed invention share a common special technical feature to have unity of invention, and that the special technical feature is not taught by the prior art. Lambowitz teaches nucleic acids comprising an intron endonuclease cleavage site, and the combination of Klickstein et al (a) and Klickstein et al (b) teach the claimed methods of inserting DNA into a vector. Therefore, Groups 1-3 do not share a special technical feature.